

Dutch Parallel Corpus User Manual

CONTENTS

1. Introduction	2
2. Corpus Design	3
2.1 Balance	3
2.2 Text types	3
2.3 Text providers	5
2.4 Copyright clearance	5
2.5 Metadata	6
2.6 Conclusion	6
3 Data processing	8
3.1 Quality control.....	8
3.2 Text normalization	8
3.3 Sentence alignment	8
3.4 Sub-sentential alignment.....	9
3.5 Linguistic annotation	10
3.5.1 Tokenization.....	10
3.5.2 Lemmatization and Part-of-Speech tagging	11
3.5.3 Syntactic information.....	12
3.6 Terminology Extraction.....	13
4 Exploitation	15
4.1 Monolingual vs. Bilingual search	15
4.2 Full corpus vs. Subcorpus	15
4.3 The search proper.....	16
4.3.1 Web interface functionality: survey	16
4.3.2 Web interface functionality: example	16
Appendices	20
Appendix one: Design in numbers	20
Appendix two: Metadata	22
Appendix three: PoS tags.....	25

1. Introduction

The present manual describes the **Dutch Parallel Corpus (DPC)**: a 10-million-word, high-quality, sentence-aligned parallel corpus for the language pairs Dutch-English and Dutch-French, with Dutch as the central language.

It contains a detailed description of the design principles underlying DPC and the different stages of data-processing. The web interface is also discussed and illustrated with various examples.

The most recent version of this text is available online at the following address: <http://www.kuleuven-kortrijk.be/dpc/manual>

Acknowledgements

DPC is a research project that was financed by the Nederlandse Taalunie (Dutch Language Union) within the framework of the STEVIN-programme¹, (a Dutch acronym for Essential Speech and Language Technology Resources), a multi-year programme stimulating research in Dutch language and speech technology.

DPC was created by a Flemish consortium (K.U.Leuven Campus Kortrijk and Faculty of Translation Studies of the University College Ghent). The core team was assisted by a number of research partners with expertise in different domains: data driven machine-learning tools, linguistic annotation, alignment and corpus exploitation. The core team collaborated closely with other Stevin projects: D-Coi, SoNaR and Lassy.

To make sure that the corpus fulfils the need of the different intended users, a user group was composed representing specialists of the different application and research domains. The user group consists of industrial and academic partners.

The following researchers contributed to DPC: Piet Desmet, Willy Vandeweghe, Hans Paulussen, Lieve Macken, Maribel Montero Perez, Orphée De Clercq, Lidia Rura, Julia Trushkina and Antoine Besnehard

¹ <http://taalunieversum.org/taal/technologie/stevin/>

2. Corpus Design

DPC consists of two language pairs: **Dutch-English** and **Dutch-French** and is bi-directional. A part of the corpus is **trilingual**.

The design principles were based on research into standards for other parallel corpus projects, and a user requirements study. Two objectives were of paramount importance: balance and quality.

In this chapter we discuss the aspect of balancing the corpus, while the next chapter focuses on how high quality was ensured for each processing step.

2.1 Balance

The DPC corpus consists of two language pairs (Dutch-English and Dutch-French), hence its four translation directions (Dutch into English, English into Dutch, Dutch into French and French into Dutch) and five text types:

- Literature
- Journalistic texts
- Instructive texts
- Administrative texts
- External communication

The corpus is balanced proportionally with respect to translation direction and text type. In accordance with DPC's design principles, the corpus contains five text types that each account for 2,000,000 words. Within each text type, each translation direction contains 500,000 words.

When constructing DPC, two exceptions were made to the global design:

- Given the difficulty to find information on translation direction for instructive texts, the condition on translation direction was relaxed for this text type.
- For literary texts, it often proved difficult to obtain copyright clearance. Due to time constraints, the literary texts are not strictly balanced according to translation direction, but are balanced according to language pair.

The exact number of words can be found in Appendix One.

2.2 Text types

In order to enhance the navigability of the corpus, a subclassification was imposed on the five text types resulting in the creation of a finer tree-like structure within each type. This subdivision has no implications on the balancing of the corpus. The introduction of subtypes is merely a way of mapping the actual landscape within each text type, and assigning accurate labels to the data in order to enable the user to correctly select documents and search the corpus.

The labels for the subtypes were chosen from cognitively tangible categories, most of them are encountered in everyday use.

The following subtypes have been distinguished within the DPC typology:

- **Literature** is subdivided into fictional and non-fictional texts. The fictional texts was not further subdivided in basic-level categories since we only manages to clear some **novels** from copyrights, which is a well-known and recognized genre. Whereas the **non-**

fictional literature is an umbrella category uniting **three** basic-level categories, all of them are well-known genres: **essays**, **(auto)biographies** and **expository works of a general nature**.

- **Journalistic texts** were roughly subdivided into **two** basic-level categories: **news reporting articles** and **comment articles**. The latter comprises background articles, columns and editorials.

- **Instructive texts** contain **three** basic-level categories: **manuals**, **legal documents** (e.g. contracts, conditions, regulations etc) and **procedural descriptions**, i.e. documents dealing with all kinds of procedures.

- **Administrative texts** comprise **five** basic-level categories: **legislation** (written law), **proceedings of parliamentary debates**, **minutes of meetings**, **yearly reports** and **official speeches**. These texts are produced within an institutional context, their circulation is usually restricted to internal use or to use within a limited circle of organizations tied to the institution.

- **External communication** consists of **five** basic-level categories: **(self-) presentations of organisations, projects, events; informative documents of a general nature; promotion and advertising material; press releases & newsletters** and **scientific texts**. These are texts of an informative and/or persuasive nature that are characterized by a wide circulation and meant for external use in general or for peers in a broad sense.

The DPC typology is presented in the table below. The five main types represent superordinates each containing several basic-level categories.

[SUPERORDINATE]		[BASIC LEVEL]
1. Literature	Fictional	1.1 Novels
	Non-fictional	1.2 Essayistic texts
		1.3 (Auto)biographies
		1.4 Expository works of a general nature
2. Journalistic texts		2.1 News reporting articles
		2.2 Comment articles (background articles, columns, editorials)
3. Instructive texts		3.1 Manuals
		3.2 Internal legal documents
		3.3 Procedure descriptions
4. Administrative texts		4.1 Legislation
		4.2 Proceedings of parliamentary debates
		4.3 Minutes of meetings
		4.4 Yearly reports
		4.5 Official speeches
5. External communication		5.1 (Self-)presentation of organisations, events
		5.2 Informative documents of a general nature
		5.3 Promotion and advertising material
		5.4 Press releases and newsletters
		5.5 Scientific texts

Table 1: Text types and subtypes included in DPC

All this information is also stored in the metadata (see 2.5).

The exact number of words per text type and translation direction can be found in Appendix One.

2.3 Text providers

To guarantee the quality of the text samples, most of them were taken from published materials or from companies or institutions working with a professional translation division. Care was taken to differentiate kinds of data providers, among them providers from publishing houses, press, government, corporate enterprises, European institutions, etc.

Differentiation was also compulsory at cell level: the material of each cell (i.e. the unique combination of text type and translation direction) originates from at least three different providers in order to preserve good balance. This is why it was decided to limit the number of words per text provider to 166,666 for every combination.

In some cases, however, this ceiling could not be respected for pragmatic reasons, and more material came from a single provider. This is the case, for example, with journalistic texts. Though it may be that articles only came from one text provider, they were in fact written or translated by various people.

2.4 Copyright clearance

In order to make the corpus accessible to the entire research community, **copyright clearance** was obtained for all samples included in the corpus. These licence agreements guarantee accessibility and protect the intellectual and economic property rights of the authors and publishers.

Four types of agreements were used: **IPR for commercial use**, **IPR for publishers**, **IPR short version** and an **e-mail or letter with permission**. All this information is stored in the metadata so as to guide each user in knowing which text material is accessible to the entire research community and which material has limited use.

In the following table we list some text providers that contributed to the DPC project for each text type and translation direction. This is not an exhaustive list, because some text providers desired to remain unknown. In total, 55 text providers participated in DPC.

Transl	Text Provider	Data
Literature (fictional & non-fictional)		
En -> DU	Little Brown -> Nijgh & Van Ditmar	Extracts of Novels (fictional)
Fr -> DU	Editions du Seuil -> Nijgh & Van Ditmar	Extracts of Novels (fictional)
Du -> En	Mercatorfonds	Expository works (non-fictional)
Du -> FR	Nijgh & Van Ditmar -> Editions du Seuil	Extracts of Novels (fictional)
Journalistic texts		
En -> DU	The Independent -> De Standaard	News Articles, Comment articles, Columns, Editorials
Fr -> DU	Roularta	News Articles, Columns
Du -> En	Campuskrant	Comment articles
Du -> FR	Roularta	News Articles, Columns

Instructive texts		
Unknown	IBM	Manuals
	Bosch	Manuals
Administrative texts		
En -> DU	Melexis	Yearly Reports
Fr -> DU	RIZIV	Yearly reports, minutes of meetings
Du -> En	Vlaamse Overheid	Yearly Reports, minutes
Du -> FR	FOD Sociale Zekerheid	Yearly reports, correspondence
External Communication		
En -> DU	Ablynx	Press releases
Fr -> DU	NMBS	Press releases
Du -> En	Arcelor Mittal	Promotion and Advertising material
Du -> FR	Transmed	Informative documents of a general nature

Table 2: Selection of text providers

2.5 Metadata

All the text material included in the corpus is annotated with additional metadata at different levels. This allows the user to retrieve relevant information from the corpus. The DPC-metadata are of two kinds: (1) **text-related** data and (2) **translation-related** data. Finally some **statistics** are added.

(1) The first kind includes information on the **text**: language, author and/or translator, publishing information, intended outcome of the text. The text is also characterized according to its type and domains, as well as according to the type of institution that produced the text (profit vs. non-profit) and according to its intended audience (internal communication, external communication for specialists, or external communication for a general public). A list of relevant keywords is provided for the text, as well as information on copyright.

(2) The second kind – **translation**-related data – indicates the translation direction, and links original and translated texts. It also notes how the text was translated (human translation, translation by a human using translation memory or machine translation corrected by a human).

The **statistics** mention how many words and sentences a certain document contains.

These metadata suit different types of users. Any user can select – according to his or her needs – a more fine-tuned sample set based on the combination of metadata tags.

More information can be found in Chapter 4, where the importance of the metadata as a first step in corpus search is underlined. For a complete list of all possible metadata tags, see Appendix Two.

2.6 Conclusion

This chapter discussed DPC's corpus design, marked by two features: balanced composition and research availability.

The Dutch Parallel Corpus contains texts from a wide range of text types and diverse domains. It contains two bidirectional bilingual parts and one trilingual part. For exact numbers of the total amount of words that are included in DPC, the user can consult Appendix One.

In order to maximize research potential, copyright clearance was obtained for all texts. DPC is made available to the research community through the Dutch Agency for Human Language Technologies (the TST-centrale).

The next chapter will expand on quality control and other data processing steps.

3 Data processing

3.1 Quality control

Since one of the explicit objectives of DPC was obtaining high quality, a quality control system was put into place for each step in compiling, aligning and annotating the corpus.

Three forms of quality control were envisaged: manual verification, a spot-checking module and automatic control.

Manual verification, traditionally the best guarantee for high quality data, was performed by qualified linguists with native and near-native language proficiency. Ten percent of the whole corpus was manually verified for each processing step (= 1 million words). The exact composition of this manually verified 1 million word corpus can be found on <http://www.kuleuven-kortrijk.be/dpc/xtra/G1/G1.html>.

The second step was to develop a **spot-checking** module on the basis of error analysis of the manually verified data. This was only done for those processing steps of which the output could be upgraded considerably using simple spot-checking heuristics.

Finally, other data processing steps were verified with **automatic control procedures**.

Each step of the data processing will now be discussed in more detail.

3.2 Text normalization

The data acquired came in different formats and thus needed to be brought into conformity with a DPC standard. To this end, every text was converted into txt-format, assigned a unique DPC name and grouped together according to text type. Graphs, tables, tables of contents and figures were removed from the material so as to end up with clean text. Text material that had originally been drawn up in PDF format required particular attention.

Considerable time and effort was devoted to this process of 'cleaning' incoming texts, to ensure that the following - more automated - steps could be carried out as smoothly as possible. All texts, i.e. 100% of the corpus, were therefore manually cleaned.

Once the text had been cleaned, it was split into sentences, a necessary step to be able to perform the next processing steps.

Quality control principles were applied for all of these steps: a manual check for 10% of the corpus, spot-checking heuristics or automatic control procedures for the remaining 90%

3.3 Sentence alignment

The whole corpus is aligned on sentence level, which means that each sentence of a source language text was linked to its target text equivalent. The sentences linked by the alignment procedure thus represent translations of each other in different languages.

The alignment procedure resulted in matches of a different kind:

- **1:1** (one sentence in a source language is aligned with one sentence in a target language)
- **1:many** (one sentence in a source language is aligned with two or more sentences in a target language)
- **many:1** (two or more sentences in a source language are aligned with one sentence in a target language)
- **many:many** (two or more sentences in a source language are aligned with two or more sentences in a target language)
- **0:1** (no alignment links for a sentence in a target language)
- **1:0** (no alignment links for a sentence in a source language)

Zero alignments were only accepted if no translation could be found for a sentence in either the source or the target language, in other words when a corresponding part of text was missing in the other language.

Many-to-many alignments were legitimate in two cases: overlapping alignments and crossing alignments. In other cases, smaller links were used. For example, unless 2:2 alignment is a true case of an overlapping or a crossing alignment, two 1:1 links were used.

An overlapping alignment is due to asymmetric sentence splitting in the two languages, whereas a crossing alignment means that the translation of a sentence in the source text shows up at another place in the target text. These two alignment types were inadmissible in DPC and therefore put under the umbrella of many-to-many alignments.

Ten percent of the sentence-aligned data was checked manually. For this manual verification the sentences were run through the Vanilla aligner. Because this aligner requires paragraph-aligned data, 10% of the corpus was also manually checked on paragraph level.

Afterwards, this manually verified output was compared with the combined output of three aligners, namely the Vanilla aligner, the Microsoft aligner and the GMA aligner, so as to be able to retrain the tools and to work out spot-check heuristics. 90% of the sentence-aligned data was verified using spot-checks.

3.4 Sub-sentential alignment

For more than 25,000 words of the Dutch-English part of the corpus, manual alignments at the sub-sentential level were created. Reference corpora where sub-sentential translational correspondences are indicated manually – also called Gold Standards – are used as an objective means for testing word alignment systems.

The reference corpus consists of journalistic texts, newsletters and medical European Public Assessment Reports. We assume that for each of the three text types another translation style was adopted. Table 4 summarizes the formal characteristics of the corpus: total number of words, average sentence length of source and target sentences

and the ratio of source-target sentences. In total, the Gold Standard contains more than 25,000 words.

Text type	Total words	Avg sentence length (source)	Avg sentence length (target)
Journalistic texts	7,706	22.0	22.0
Newsletters	10,480	15.0	15.4
EPARs ²	7,536	17.2	17.7

Table 3: Sub-sententially aligned corpus

To account for a wide range of translational phenomena, three types of links were introduced: regular links are used to connect straightforward correspondences; fuzzy links for translation-specific shifts of various kinds (paraphrases and divergent translations); and null links for source text units that have not been translated or target text units that have been added.

A multi-level annotation is proposed in case of divergent translations: fuzzy links are used to connect paraphrased sections, regular links are used to connect corresponding words within the paraphrased sections.

The annotation guidelines are available on the website of LT3 (<http://veto.hogent.be/lt3/>). For more information we refer to:

Lieve Macken (2010) An annotation scheme and Gold Standard for Dutch-English word alignment. Proceedings of the Seventh International Conference on Linguistics Resources and Evaluation (LREC-2010), Valletta, Malta.

3.5 Linguistic annotation

Linguistic annotation involves lemmatization and part-of-speech (PoS) tagging of the DPC data, two processing steps which are usually linked together. The input data had to be tokenized, a pre-processing task which is performed before the actual tagging procedure.

The whole corpus is tokenized, lemmatized and enriched with Part-of-Speech tags. Since these steps are language dependent, different tokenizers, lemmatizers and PoS-taggers were used for every DPC-language.

3.5.1 Tokenization

During tokenization a sentence is split into sequences of words. All punctuation marks not belonging to the word form (i.e. punctuation marks that are not part of an abbreviation) are stripped off. Differences between the tokenization procedures for all three languages are related to the tagging tools used.

The treatment of certain punctuation marks required a different approach depending on the language. An example is the treatment of the possessive marker 's in English and Dutch. According to the conventions of the English part-of-speech taggers, the possessive marker 's is split off during tokenization, and a separate PoS tag is assigned. The conventions of the Dutch part-of-speech tagger, on the other hand, do not bring about the possessive marker to be stripped off during tokenization, as possessiveness of the noun is coded in the PoS tag.

² EPAR stands for European Public Assessment Report, this text type includes patient information leaflets

Tokenization for Dutch was performed by the D-COI tokenizer, for English a slightly adapted version was used. The French data was tokenized with the help of an adapted version of the French tokenizer scripts, which is part of the TreeTagger program documentation.

After the linguists had manually checked 10% of the output, the development of spot-checking modules to tokenize the remaining 90% of the corpus, turned out to be superfluous as the tokenizers could proceed automatically, given slight adaptations.

3.5.2 Lemmatization and Part-of-Speech tagging

The lemmatization process generates the base form (lemma) for each orthographic token. Part-of-speech tagging assigns a part-of-speech code to each orthographic token.

The lemmatizers for the three languages use similar definitions of base form or lemma. The base form for verbs is the infinitive, with other words it is the stem, i.e. the word form without inflectional affixes.

Although ideally one would like to compare grammatical codes over the three languages, limitations of the tools and inherent features of the three languages involved, do not allow for a straightforward mutual mapping of the PoS codes. It was decided to use widely accepted PoS tag sets for each language. In Appendix Three you can find the frequency of the head PoS tags for each language.

Dutch

The PoS tagging system and tools developed within D-COI (a corpus project for Dutch) were borrowed for the Dutch section of DPC. The advantage of this is that the Dutch data – Dutch being the central language in DPC – can be directly related to existing Dutch corpora, thus allowing for transparent search queries in linked Dutch corpora, whenever need be.

For the 1 million subcorpus, the ensemble tagger was used.

The ensemble tagger uses the CGN PoS tagset (Van Eynde, Frank and Zavrel, Jakub and Daelemans, Walter 2000), which is characterized by a high level of granularity. Apart from the word class, the CGN PoS tag set codes a wide range of morphosyntactic features as attributes to the word class. In total, 316 distinct full tags are discerned.

The D-COI procedures were observed for the 10% manual verification of the Dutch PoS tags and the lemmata, the procedures of the D-COI project were used. This implies that only the words for which the different taggers do not agree were manually verified. The DCOI-protocol³ with its description of all the possible tags served as a reference guide. In addition to the verification of PoS and lemma, we also grouped multiword units and Dutch separable verbs, using the CGN protocol⁴ as a reference.

Lemmatization and PoS tagging of the 9M corpus was also effectuated with the help of the ensemble tagger. The tagging task was carried out by the team of ILK Research group Tilburg.

³ http://www.ccl.kuleuven.be/Papers/POSmanual_febr2004.pdf

⁴ http://lands.let.kun.nl/cgn/doc_Dutch/topics/version_1.0/annot/lex_linkup/lxk_prot.pdf

English

For English, part-of-speech tagging and lemmatization for English was performed by the combined memory-based PoS tagger/lemmatizer, which is part of the MBSP tools (Daelemans, Walter and Buchholz, Sabine and Veenstra, Jorn 1999) and (Daelemans, Walter and Van den Bosch, Antal 2005). The English memory-based tagger was trained on data from the Wall Street Journal corpus in the Penn Treebank (Marcus, Mitchell P. and Santorini, Beatrice and Marcinkiewicz, Mary Ann 1993), and uses the Penn Treebank tagset. The Penn Treebank tagset contains 45 distinct tags.

All PoS codes and lemmata of 10% of the corpus were manually inspected and verified. For this we used the PennTreebank Tagging guidelines⁵ as a reference.

These manually verified annotations were used to test the performance of two different PoS taggers using the same PoS tag set and lemmatisation conventions: the MBSP tagger and Treetagger. As both taggers made different errors, we combined the output of both taggers to process the nine million word corpus and only verified the PoS tags and lemmata for which both taggers did not agree. With a limited manual verification effort, we can achieve 98% precision⁶ for PoS tagging and 99% precision for lemmatization.

French

In order to manually check 10% of the data, the linguistic annotation was done by using the combined output of the French version of TreeTagger. In fact, the first run used the tag set of the original TreeTagger and FLEMM lemmatisation information. In the second run, the LIMSI tagset was used. The output of both tagging procedures were compared during the analysis of the 1M set, and a quality procedure was developed in order to spot check the data from the 9M set.

The tagset consists of 312 morphosyntactic tags.

Allauzen, Alexandre and Hélène Bonneau-Maynard (2008), "Training and Evaluation of POS Taggers on the French MULTITAG Corpus". In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), pages 28-30.

Paroubek, Patrick (2000), "Language resources as by-product of evaluation: the multitag example". In Second International Conference on Language Resources and Evaluation (LREC) 2000, pages 151-154.

3.5.3 Syntactic information

A smaller part of the DPC data is syntactically annotated.

The Dutch selection (200,000 words) was annotated by the LASSY team who used the Alpino parser (developed at Groningen University) for this.

The texts were selected from the following 4 text types:

- administrative texts: 26,3520 words
- instructive texts: 25,985 words
- external communication: 66,379 words
- journalistic texts: 81,104 words

⁵ <http://www.inf.unibz.it/~bernardi/Courses/CompLing/Papers/tagguide.pdf>

⁶ estimations derived from the 1 million word corpus

The texts come from both the Dutch/French and the Dutch/English part of the corpus and contain texts originally written in Dutch as well as texts translated into Dutch.

3.6 Terminology Extraction

In order to evaluate different terminology extraction tools, a **Gold Standard** (i.e. a manually created reference set) for terminology extraction was created within the framework of the DPC-project. Terminology extraction can be seen as a first step towards terminology management. In the terminology extraction phase, terms are identified in a text and – in the case of multilingual terminology extraction – the corresponding translations are retrieved. The extracted terms and their translations can be stored in bilingual glossaries, which are already a valuable aid for technical translators. If the aim is the creation of a term bank, the extracted terms are structured in concept-oriented databases in the terminology management phase.

The Gold Standard contains texts of two different domains:

- **Medical domain:** trilingual texts (Dutch/French/English)
- **Financial domain:** bilingual texts (Dutch/French and Dutch/English)

In the Gold Standard, all terms (single- and multiword terms) were manually indicated. As we had no domain experts to our disposal, all terms were looked up in several reference books. The reference books in which the terms were found are included in data set.

Details on the texts of the extraction corpus are presented in the following table.

Domain	Texts	Lang Pairs	Dutch words	Terms
Financial	ELI	DU/EN/FR	11,365	469
Medical	ING	DU/EN	9,458	400
	QTY	DU/FR	8,954	338

Table 4: Extracted terms for the gold standard

The following texts were included in the extraction corpus:

ELI	ING	QTY
dpc-eli-000937	dpc-ing-001878	dpc-qty-000928
dpc-eli-000938	dpc-ing-001879	dpc-qty-000930
dpc-eli-000939	dpc-ing-001888	dpc-qty-000932
dpc-eli-000940		dpc-qty-000933
dpc-eli-000941		dpc-qty-000935
dpc-eli-000942		dpc-qty-000936 ⁷
dpc-eli-000943		
dpc-eli-000944		
dpc-eli-000945		
dpc-eli-000946		
dpc-eli-000947		
dpc-eli-000948		

⁷ The text of dpc-qty-000936 was shortened (the sections from "U bent wijnbouwer in Waals-Brabant." until "Ik denk dat ik niet zo'n slecht tacticus ben want ik handel snel en doeltreffend en heb een goed zicht op wat er binnen vijf jaar op het spel staat." were deleted because they contained text not dealing with the financial domain).

The resulting term lists consists of three or four fields delimited by a tab. The first field contains the Dutch term, the last field contains reference codes; the one or two other fields contain the translations.

An overview of the codes to the reference books is given below:

CODE	SOURCE	ONLINE
AZWIKI	Gezondheid van A-Z Wikipedia	http://nl.wikipedia.org/wiki/Gezondheid_van_A_tot_Z
DGF	Dictionnaire de la comptabilité et de la gestion financière, Louis	-
DICM	Dictionnaire Médical, Manuila, Lewalle, Nicoulin	-
DMF	Dictionnaire Médical Flammarion	-
DMFI	Dictionnaire des marchés financiers, Antoine & Capiou-Huart	-
DTM	Dictionnaire français des termes de médecine	-
ELSE	Elsevier's dictionary of financial terms (English, French et al.), Marie-Claude Bignaud	-
EUR	Euramis, terminologiebank Europese Commissie	-
EURFR	Eureka Santé	http://www.eurekasante.fr/lexique-medical.html
FELNE	Financieel Economisch Lexicon N-E, A.J. de Keizer	-
FINCAN	Glossaire Financier (Canadees)	http://www.lautorite.qc.ca/userfiles/File/Publications/Consommateurs/Glossaire.pdf
IDF	International Dictionary of Finance, The Economist Books	http://www.medicalreflex.fr/grand-public/
LEXMED	Lexique de terminologie médicale	http://georges.dolisi.free.fr/Terminologie/Menu/terminologie_medicale_menu.htm
MEDREF	Medical Reflex	-
MEDSAN	Lexique Médical Médecine et Santé	http://www.medecine-et-sante.com/lexique.html
MWENNE	Medisch woordenboek E-N/N-E, Mostert	-
PINK	Pinkhof Geneeskundig Woordenboek	-
TAALVL	Taalvlinder	http://www.taalvlinder.com/pages/medici.htm
UBSLEXFR	Lexique bancaire UBS	http://www.ubs.com/1/f/about/bterms.html
WGENNE	Woordenboek geneeskunde E-N/N-E, Kerkhof	-
ZIEK	Ziekenhuis.nl	http://www.ziekenhuis.nl/index.php

Table 5: Consulted reference works

4 Exploitation

The corpus can be exploited either as a full text resource or as a web search interface. This chapter focuses on the web interface that was developed for the different users of the corpus doing a corpus search.

The web interface was developed by Geert Peeters & Serge Verlinde (ILT, Leuven), it was composed for users with notions of Dutch.

4.1 Monolingual vs. Bilingual search

The first choice an intended user can make is whether he/she wants to make a monolingual or bilingual search.

Monolingual:

The screenshot shows the 'DPC SEARCH' web interface. At the top, it says 'DPC SEARCH' and 'Klik hier voor de handleiding'. Below this, there are two main sections: 'Ééntalig zoeken' (Monolingual search) and 'Tweetalig zoeken' (Bilingual search). In the 'Ééntalig zoeken' section, there is a dropdown menu labeled 'taal:' with the option 'make your choice' selected. The dropdown menu is open, showing a list of language options: EN, EN (UK), EN (US), FR, FR (BE), FR (FR), NL, NL (BE), and NL (NL). In the 'Tweetalig zoeken' section, there are two dropdown menus labeled 'taal1:' and 'taal2:'. 'taal1:' has 'EN' selected, and 'taal2:' has 'NL' selected. Below these sections, there are radio buttons for 'doorzoek' (search) and 'alle teksten' (all texts), with 'alle teksten' being selected. There is also a 'Verstuur' (Send) button.

Bilingual:

The screenshot shows the 'DPC SEARCH' web interface. At the top, it says 'DPC SEARCH' and 'Klik hier voor de handleiding'. Below this, there are two main sections: 'Ééntalig zoeken' (Monolingual search) and 'Tweetalig zoeken' (Bilingual search). In the 'Ééntalig zoeken' section, there is a dropdown menu labeled 'taal:' with the option 'make your choice' selected. In the 'Tweetalig zoeken' section, there are two dropdown menus labeled 'taal1:' and 'taal2:'. 'taal1:' has 'EN' selected, and 'taal2:' has 'NL' selected. The dropdown menu for 'taal2:' is open, showing a list of language options: EN, EN (UK), EN (US), FR, FR (BE), FR (FR), NL, NL (BE), and NL (NL). Below these sections, there are radio buttons for 'doorzoek' (search) and 'alle teksten' (all texts), with 'alle teksten' being selected. There is also a 'Verstuur' (Send) button.

4.2 Full corpus vs. Subcorpus

Secondly the user has the choice whether he would like to search the entire corpus or a smaller part, a subcorpus.

A subcorpus can be put together by using the DPC metadata which exhibit a whole range of features that allow the user to make a number of selections on the DPC web interface. A user can specify, for instance, what types of texts, languages and domains should be used in the search. For a complete list of all different metadata, please consult Appendix Two.

Putting together a subcorpus:

The screenshot shows the 'DPC SEARCH' web interface. At the top, there's a header with the title 'DPC SEARCH' and a sub-header 'Klik hier voor de handleiding'. Below this, there are two main sections: 'Eéntalig zoeken' (Monolingual search) and 'Tweetalig zoeken' (Bilingual search). The 'Eéntalig zoeken' section has a 'taal:' dropdown menu with the option 'make your choice'. The 'Tweetalig zoeken' section has two dropdown menus: 'taal1:' with options 'EN' and 'en', and 'taal2:' with options 'NL' and 'en'. Below these sections, there are radio buttons for 'doorzoek' (selected), 'alle teksten', and 'tekstselectie'. A 'subcorpus' section contains several dropdown menus: 'text type' (with a list including 'External Communication', 'Fictional literature', 'Instructive texts', 'Journalistic texts', and 'Non-fictional literature'), 'domain' (with 'make your choice'), 'copyright / ip-agreement' (with 'make your choice'), 'original text & language' (with 'make your choice'), and 'translated text & language' (with 'make your choice'). There are also two lines of text: 'Het subcorpus van taal 1 bestaat uit 540820 woorden.' and 'Het subcorpus van taal 2 bestaat uit 997358 woorden.' A 'Verstuur' button is located at the bottom right of the 'subcorpus' section.

On the basis of the selected metadata, the user creates a new corpus that can be presented as a DPC subcorpus. This subcorpus constitutes the starting point for any further search on the web interface.

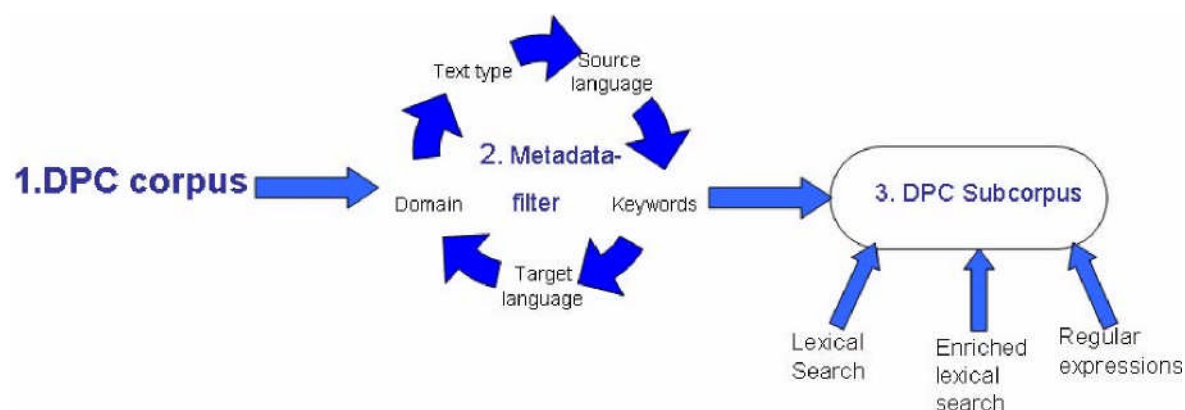
The metadata can thus be defined as a first filter in the search task on the interface.

4.3 The search proper

4.3.1 Web interface functionality: survey

After selecting the metadata and putting up a subcorpus, the user can execute a second search command for the research proper.

This process is visualized in the following graph.



4.3.2 Web interface functionality: example

Step 1: Select a subcorpus

In this example we carry out a bilingual search on a selection of texts (Administrative texts with French as original language and Dutch as translated language).

DPC SEARCH
Klik hier voor de handleiding

Eéntalig zoeken

taal: make your choice

Tweetalig zoeken

taal1: FR en taal2: NL

doorzoek ☐ alle teksten - ☒ tekstselectie

subcorpus

text type: Administrative texts

domain: make your choice

copyright / ipr-agreement: make your choice

original text & language: FR

translated text & language: NL

Het subcorpus van taal 1 bestaat uit 283752 woorden.
Het subcorpus van taal 2 bestaat uit 276066 woorden.

Verstuur

We see that our subcorpus for French contains 283,752 words and the one for Dutch 276,066 words.

Step 2: Define your search query

Once the subcorpus has been determined, the user can perform his search. In our example we carry out an enriched bilingual search for the language pair Dutch-French. The example given below concerns the use of past tenses in French and Dutch. In our bilingual search, we search examples that contain the French verb “avoir” used as an auxiliary in “indicatif présent” and that is followed by a past participle. As for the Dutch component, we select verbs on lemma that are used in the past tense. Besides, we also specify that the results for Dutch cannot contain a past participle. Our search thus excludes the literal translation of the French “passé composé”.

nieuwe zoekopdracht

subcorpus

text type: Administrative texts

domain: make your choice

copyright / ipr-agreement: make your choice

original text & language: FR

translated text & language: NL

tweetalig zoeken Frans - Nederlands

woord1

avoir op ☐ woordvorm ☒ lemma (* is jokerteken, ? is één karakter jokerteken)

parts of speech

verbe ☐ tous les verbes ☐ verbe principal ☒ verbe auxiliaire [sluiten]

indicatif

☒ sing. + plur. ☐ singulier ☐ pluriel

☒ 1e + 2e + 3e pers. ☐ 1e pers. ☐ 2e pers. ☐ 3e pers.

☐ tous les temps ☒ présent ☐ imparfait ☐ futur simple ☐ passé simple

positie

de positie is niet belangrijk

☒ EN ☐ NIET

woord2

op ☐ woordvorm ☒ lemma (* is jokerteken, ? is één karakter jokerteken)

parts of speech

verbe ☐ tous les verbes ☒ verbe principal ☐ verbe auxiliaire [sluiten]

Maak uw keuze

positie

de positie is niet belangrijk

maximum aantal woorden tussen de zoekwoorden: 0

alles wissen

woord1

op ☒ woordvorm ☐ lemma (* is jokerteken, ? is één karakter jokerteken)

parts of speech

werkwoord ☐ alle vormen ☐ enkelvoud ☐ meervoud ☐ met -t [slu]

persoonsvorm

☒ alle tijden ☐ tegenwoordige tijd ☒ verleden tijd ☐ conjunctief

positie

de positie is niet belangrijk

☐ EN ☒ NIET

woord2

by ☐ woordvorm ☒ lemma (* is jokerteken, ? is één karakter jokerteken)

parts of speech

werkwoord ☐ voltooid deelwoord [slu]

positie

de positie is niet belangrijk

maximum aantal woorden tussen de zoekwoorden:

toon 100 resultaten zoek

Step 3: Results

The results are shown under your search specifications and the words are indicated in red. These results can also be exported to an excel sheet.

Aantal resultaten: 21 waarvan er 21 getoond worden.

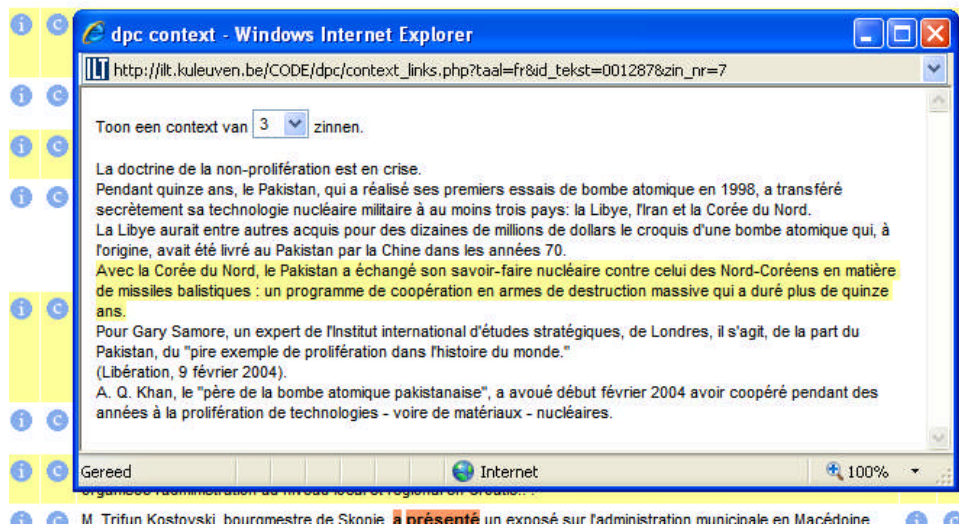
 Exporteer deze resultaten naar Excel.

 	Avec la Corée du Nord, le Pakistan a échangé son savoir-faire nucléaire contre celui des Nord-Coréens en matière de missiles balistiques : un programme de coopération en armes de destruction massive qui a duré plus de quinze ans.	 	Met Noord-Korea ruilde Pakistan zijn eigen nucleaire knowhow voor de Koreaanse knowhow inzake ballistische raketten. Zo ontstond met betrekking tot massavernietigingswapens een samenwerkingsprogramma dat meer vijftien jaar duurde .
 	Le président de l'AP OSCE, M. Goran Lenmarker, a souligné que la conférence s'intéresserait principalement à la situation des pays de l'Europe du Sud-Est...	 	De voorzitter van de PA OVSE, de heer Goran Lenmarker, beklemtoonde dat deze conferentie vooral oog h ebben voor de situatie in Zuidoost Europa.
 	I a insisté sur l'importance du dialogue interparlementaire dans la recherche de solutions durables pour assurer la stabilité dans la région.	 	Hij onderstreepte het belang van de interparlementaire dialoog in de zoektocht naar lange termijn oplossingen stabiliteit in deze regio.
 	Au nom de M. Miguel Angel Moratinos, président en exercice de l'OSCE et ministre espagnol des Affaires étrangères, M. Josep Borell, représentant spécial du président en exercice, a attiré l'attention sur le rôle important de l'assemblée parlementaire de l'OSCE dans le renforcement de l'état de droit démocratique... Il a dressé un bilan des dernières évolutions dans la zone OSCE et exposé les priorités de la présidence espagnole.	 	Namens de heer Miguel Angel Moratinos, de Chairman-in-Office van de OVSE en de Spaanse minister van Buitenlandse Zaken, wees de heer Josep Borell, speciaal gezant van de Chairman-in-office, op de belangrijke van de parlementaire assemblee van de OVSE bij de versterking van de democratische rechtsstaat en ga hij e overzicht van de actuele ontwikkelingen in de OVSE-zone en de aandachtspunten van het Spaanse voorzitterschap.
 	Au nom de M. Miguel Angel Moratinos, président en exercice de l'OSCE et ministre espagnol des Affaires étrangères, M. Josep Borell, représentant spécial du président en exercice, a attiré l'attention sur le rôle important de l'assemblée parlementaire de l'OSCE dans le renforcement de l'état de droit démocratique. Il a dressé un bilan des dernières évolutions dans la zone OSCE et exposé les priorités de la présidence espagnole.	 	Namens de heer Miguel Angel Moratinos, de Chairman-in-Office van de OVSE en de Spaanse minister van Buitenlandse Zaken, wees de heer Josep Borell, speciaal gezant van de Chairman-in-office, op de belangrijke van de parlementaire assemblee van de OVSE bij de versterking van de democratische rechtsstaat en ga hij e overzicht van de actuele ontwikkelingen in de OVSE-zone en de aandachtspunten van het Spaanse voorzitterschap.
 	M. Borell a souligné que les représentations de l'OSCE (les "field missions") contribuent de manière importante à assurer la sécurité régionale...	 	De heer Borell wees op de belangrijke bijdrage van de OVSE vertegenwoordigingen (de "field missions") om c veiligheid in de regio's te verzekeren.
 	M. Antun Palarić, secrétaire d'État du Bureau Central pour l'Administration en Croatie, a expliqué comment était organisée l'administration au niveau local et régional en Croatie...	 	De heer Antun Palarić, staatssecretaris van het Centraal Bureau voor administratie in Kroatië, ga een overzicht van het bestuur op lokaal en regionaal niveau in Kroatië.
 	M. Trifun Kostovski, bourgmestre de Skopje, a présenté un exposé sur l'administration municipale en Macédoine, où un processus de décentralisation a débuté en 2001...	 	De heer Trifun Kostovski, burgemeester van Skopje, ga een uiteenzetting over het stedelijk bestuur in Macedo waar een proces van decentralisatie plaatsvond vanaf 2001.
 	M. Kostovski a insisté sur le statut spécial de Skopje, permettant à la ville de jouer son rôle de capitale et d'avoir plus facilement accès aux projets européens...	 	De heer Kostovski drong aan op een speciaal statuut voor de hoofdstad Skopje dat deze stad toelaat zijn hoofdstedelijke rol te vervullen en vlotter toegang te krijgen tot Europese projecten.
 	Lors du débat qui a suivi ces exposés, Mme Haering (Suisse) a insisté sur la nécessité de conférer un statut spécial aux capitales pour relever adéquatement les défis qui se posent.	 	het debat dat op deze uiteenzettingen volgde , wees mevrouw Haering (Zwitserland) op de nood van een spt statuut voor de hoofdsteden om de uitdagingen die zich stellen passend te beantwoorden.
 	M. Soares (Portugal) a mis l'accent sur la relation étroite entre un mandat local et un mandat parlementaire...	 	De heer Soares (Portugal) onderstreepte de nauwe verbondenheid van een lokaal met een parlementair mar
 	Le projet qui visait notamment à rassembler des enfants issus de milieux très différents a touché enfants en 2005.	 	Het project had tot doel kinderen van zeer verschillende afkomst samen te brengen en bereikte in 2005 kinde
 	M. Nery (France) a rappelé la position française...	 	De heer Nery (Frankrijk) herinnerde aan het Franse standpunt.
 	Un observateur de l'Assemblée du Conseil de l'Europe a rappelé que les règles de Bologne se greffaient sur le Processus de Lisbonne et qu'il ne fallait pas oublier le lien entre enseignement et croissance économique.	 	Een waarnemer van de Assemblée van de Raad van Europa wees erop dat de Bologneregels bij het Lissabonproces aansluiten en dat men het verband tussen onderwijs en economische groei niet mag vergeten
 	Au cours du débat qui a suivi ces exposés, M. Ivanchenko (Fédération de Russie) a indiqué qu'à horizon 2020, la Fédération russe augmenterait sa production d'énergie en associant à la sécurité régionale...	 	Spreekster wees op het belang van energie efficiëntie en hernieuwbare energiebronnen.
 	Par ailleurs, M. Soares a indiqué que le Portugal était très avancé dans le domaine de la production d'énergie solaire et éolienne...	 	energiemarkt zou zelfs de Europese Unie kunnen overstijgen en de OVSE zou hierin een rol kunnen spelen.
 	Le représentant spécial de la Région méditerranéenne, M. Alcee Hastings (USA), a évoqué les quelque 2,2 millions de réfugiés irakiens qui, à l'heure actuelle, sont principalement localisés en Jordanie, en Syrie et en Iran, et ont	 	De voorzitter van de PA OVSE, de heer Lenmarker, beklemtoonde het belang van een goed nabuurschap en gezonde economische relaties met de Mediterraan partners en een gezamenlijke aanpak inzake uitdagingen

Next to each sentence there are two little icons. When you stand on the i (information) you see the different metadata.

 	1 Filename	dpc-kam-001287	contre celui des Nord-Coréens en matière ction massive qui a duré plus de quinze
 	2 Language	FR (BE)	érence s'intéresserait principalement à la
 	3 Author/translator	Muriel Gerkens	che de solutions durables pour assurer la
 	4 Text unit title	Proposition de résolution relative à la prolifération de technologies et de matériaux nucléaires	ministre espagnol des Affaires e, a attiré l'attention sur le rôle important droit démocratique... Il a dressé un bilan présidence espagnole.
 	6 publisher	La chambre des représentants de Belgique	ministre espagnol des Affaires e, a attiré l'attention sur le rôle important de t démocratique. Il a dressé un bilan des dence espagnole.
 	8 date of publication	2007-07-18	s") contribuent de manière importante à
 	16 keywords	matériaux nucléaires	en Croatie, a expliqué comment était
 	18 Intended outcome	written to be spoken	administration municipale en Macédoine,
 	19 Text type	Administrative texts	de jouer son rôle de capitale et d'avoir
 	20 Text subtype	Legislation	sur la nécessité de conférer un statut
 	21 Domain	Foreign affairs	local et un mandat parlementaire... .
 	22 Keywords	--- Foreign affairs	ès différents a touché enfants en 2005.
 	23 Copyright/IPR-agreement	letter or e-mail with permission	
 	24 Type of institution	non-profit	
 	25 Intended audience	broad external audience	
 	26 Original text & language	FR	
 	27 Translated text & language	NL	
	28 Intermediate text & language	not applicable	
	29 Translation mode	human	
	30 Number of words	1609	
	31 Number of sentences	0	
	M. Nery (France) a rappelé la position française... .		

When clicking on the c (context) another window opens which allows you to look at the sentence in its context (ranging from 1 to 50 sentences). The sentence in yellow is the original one.



Appendices

Appendix one: Design in numbers

In accordance with the DPC design principles, the corpus is balanced in two ways: it contains five text types that each account for 2,000,000 words and each translation direction contains 500,000 words. This leads to a corpus that can be resumed in the following table. For each text type we see how many words are included per translation direction.

Text Type	SRC→TGT	DU	EN	FR	TOTAL	%
Administrative Texts	EN→DU	255,155	246,137	0	501,292	100.26
	FR→DU	307,886	0	322,438	630,324	126.06
	DU→EN	249,410	257,087	0	506,497	101.30
	DU→FR	280,584	0	301,270	581,854	116.37
	Total	1,093,035	503,224	623,708	2,219,961	111.00
External Communication	EN→DU	278,515	272,460	0	550,975	110.19
	FR→DU	233,277	0	250,604	483,881	96.78
	DU→EN	246,448	255,634	0	502,082	100.42
	DU→FR	241,323	0	270,074	511,397	102.28
	XDE-	21,679	20,118	0	41,797	8.36
	XDEF-	14,192	14,953	15,743	44,888	8.98
	Total	1,035,434	563,165	536,421	2,135,020	106.75
Instructive Texts	EN→DU	340,097	327,543	0	667,640	133.53
	FR→DU	40,487	0	42,017	82,504	16.50
	DU→EN	19,011	20,696	0	39,707	7.94
	DU→FR	110,278	0	115,034	225,312	45.06
	XD-F	59,791	0	73,758	133,549	27.71
	XDE-	299,996	296,698	0	596,694	119.34
	XDEF	138,673	145,103	166,836	450,612	90.12
	Total	1,008,333	790,040	397,645	2,196,018	109.80
Journalistic Texts	EN→DU	262,768	264,900	0	527,668	105.53
	FR→DU	240,785	0	265,530	506,315	101.26
	DU→EN	250,580	259,764	0	510,344	102.07
	DU→FR	314,989	0	340,319	655,308	131.06
	Total	1,069,122	524,664	605,849	2,199,635	109.98
Literature	EN→DU	148,488	143,185	0	291,673	58.33
	FR→DU	186,799	0	186,799	373,419	74.68
	DU→EN	346,802	361,140	0	707,942	141.59
	DU→FR	323,158	0	348,343	671,501	134.30
	Total	1,005,247	504,325	534,963	2,044,535	102.23
Grand Total		5,211,171	2,885,418	2,698,586	10,795,175	107.95

A small part of the corpus is trilingual and contains Dutch texts translated into both English and French. The following table represents the number of Dutch words that were translated in English and French per text type.

Literature	223,322
Journalistic texts	/
Instructive texts	165,205
Administrative texts	4,383
External communication	76,319
Total	469,229

Appendix two: Metadata

All text material included in DPC is provided with metadata at two levels: text-related data and translation-related data.

In the following tables a comprehensive overview of all possible metadata tags is presented.

Text-related data		
1. Language	NL(NL)	
	NL(BE)	
	EN(UK)	
	EN(US)	
	FR(FR)	
	FR(BE)	
2. Author/translator	X	
3. Text unit title	X	
4. Publishing info	magazine/journal title	
	publisher	
	ISBN/ISSN	
	date of publication	
	original date of publication	
	place of publication	
	original place of publication	
	info on previous editions	
	info on previous editions	
	editor	
	article number	
	page of the article in the magazine	
	keywords	
5. Intended outcome	class of the article	
	written to be read	
	written to be spoken	
6. Text type -> 7. Text subtype	written reproduction of spoken language	
	Literature (fictional)	Novels
	Literature (non-fictional)	Essayistic texts
		(Auto)biographies
		Expository works of a general nature
	Journalistic texts	News articles
		Comment articles (background articles)
		Comment articles (columns)
		Comment articles (editorials)
	Instructive texts	Manuals
		Internal legal documents
		Procedure descriptions
	Administrative texts	Legislation
		Proceedings of parliamentary debates
		Minutes of meetings
		Yearly reports

		Official speeches
	External Communication	(Self-)presentations of organisations, projects, events
		Informative documents of a general nature
		Promotion and advertising material
		Yearly reports
		Press releases and newsletters
		Scientific texts
8. Domain -> 9. Keywords	Communication	ICT
		Internet
	Consumption	Household appliances
	Culture	Museum
		Architecture
		Arts
		Languages
	Economy	Business
	Environment	Conservation
		Pollution
		Threats
		Nature
	Finance	Banking
		Investment
	Foreign affairs	EU
	Institutions	Management
		Policy
		Legal Documents
	Justice	Legislation
	Leisure	Tourism
		Sports
	Science	Linguistics
		Oceanography
		Zoology
		Botany
		Medicine
		Technology
	Welfare state	Social security
		Public health
		Working conditions
		Pensions
		Benifits
10. Copyright/IPR-agreement	Full version	
	Light version	
	Short version	
	Letter or e-mail with permission	
11. Type of institution	Profit	
	Non-profit	
12. Intended audience	Broad external audience	
	Limited internal audience	
	Specialist audience	

Translation-related data	
13. Original Text & Language	EN
	FR
	NL
	Unknown
14. Translated Text & Language	EN
	NL
	EN,FR
	EN,NL
	FR,NL
	EN,FR,NL
15. Intermediate Language	EN
	FR
	NL
	Unknown
	Memory
	Machine
	Unknown

Statistics	
17. Number of words	X
18. Number of sentences	X

Extra	
19. Subdocuments	X

Appendix three: PoS tags

In the following tables you see the frequency of some PoS tags in the three DPC languages: Dutch, English, French.

For Dutch and French, the PoS tags have been truncated: the subcategory labels have been stripped from the category label. In French, for example, the PoS code Ncfs (nom commun féminin singulier) has been truncated to the main category N (noun). In Dutch, for example, the PoS code VNW(aanw,adv-pron,stan,red,3,getal) has been truncated to the main category VNW (personal pronoun). In the case of English, all PoS tags are shown, with the exclusion of those tags referring to punctuation marks or other word delimiting codes.

ENGLISH		
Tag	Description	Frequency
CC	conjunction, coordinating	105,174
CD	numeral, cardinal	74,206
DT	determiner	321,793
EX	existential there	4,519
FW	foreign word	2,499
IN	preposition or conjunction, subordinating	371,991
JJ	adjective or numeral, ordinal	225,432
JJR	adjective, comparative	10,755
JJS	adjective, superlative	4,179
LS	list item marker	6,087
MD	modal auxiliary	37,866
NN	noun, common, singular or mass	516,755
NNP	noun, proper, singular	249,016
NNPS	noun, proper, plural	11,424
NNS	noun, common, plural	188,261
PDT	pre-determiner	2,089
POS	genitive marker	11,050
PRP	pronoun, personal	69,287
PRP\$	pronoun, possessive	30,951
RB	adverb	112,449
RBR	adverb, comparative	6,058
RBS	adverb, superlative	1,544
RP	particle	8,890
SYM	symbol	19,785
TO	"to" as preposition or infinitive marker	68,962
UH	interjection	937
VB	verb, base form	103,882
VBD	verb, past tense	57,842
VBG	verb, present participle or gerund	51,929
VCN	verb, past participle	92,999
VBP	verb, present tense, not 3rd person singular	51,052
VBZ	verb, present tense, 3rd person singular	77,792
WDT	WH-determiner	15,412
WP	WH-pronoun	6,505
WP\$	WH-pronoun, possessive	380
WRB	Wh-adverb	10,118
Total		2,929,870

DUTCH		
ADJ	Adjective	472,962
BW	Adverb	252,115
LET	List item	757,264
LID	Article	685,016
N	Noun (proper, common)	1,495,304
SPEC	Abbreviation, First name last name	237,959
TSW	Interjection	1,124
TW	Numeral	151,860
VG	Conjunction	294,788
VNW	Pronoun	432,513
VZ	Preposition	890,746
WW	Verb	879,826
Total		6,555,902
FRENCH		
A	Adjective	268,961
C	Conjunction	142,177
D	Determiner	507,808
F	Punctuation	425,053
I	Interjection	2,805
N	Nouns	1,019,973
P	Pronoun	169,992
R	Adverb	159,950
S	Preposition	585,868
V	Verb	417,852
X	Miscellaneous	7,746
Total		3,708,185